BIOINFORMATICS

Vol. 21 Suppl. 1 2005, pages i495–i501
doi:10.1093/bioinformatics/bti1024

Classifying noisy protein sequence data: a case study of immunoglobulin light chains

Chenggang Yu[1,3,], Nela Zavaljevski[1,3], Fred J. Stevens[1], Kelly Yackovich[2] and Jaques Reifman[3]*

[1]Argonne National Laboratory, 9700 S. Cass Avenue, Argonne, IL 60439, USA, [2]Department of Computer Information Science, Clarion University of Pennsylvania, Clarion, PA 16214, USA and [3]US Army Medical Research and Materiel Command, 504 Scott Street, Fort Detrick, MD 21702, USA

Received on January 15, 2005; accepted on March 27, 2005

ABSTRACT

Summary: The classification of protein sequences obtained from patients with various immunoglobulin-related conformational diseases may provide insight into structural correlates of pathogenicity. However, clinical data are very sparse and, in the case of antibody-related proteins, the collected sequences have large variability with only a small subset of variations relevant to the protein pathogenicity (function). On this basis, these sequences represent a model system for development of strategies to recognize the small subset of function-determining variations among the much larger number of primary structure diversifications introduced during evolution. Under such conditions, most protein classification algorithms have limited accuracy. To address this problem, we propose a support vector machine (SVM)-based classifier that combines sequence and 3D structural averaging information. Each amino acid in the sequence is represented by a set of six physicochemical properties: hydrophobicity, hydrophilicity, volume, surface area, bulkiness and refractivity. Each position in the sequence is described by the properties of the amino acid at that position and the properties of its neighbors in 3D space or in the sequence. A structure template is selected to determine neighbors in 3D space and a window size is used to determine the neighbors in the sequence. The test data consist of 209 proteins of human antibody immunoglobulin light chains, each represented by aligned sequences of 120 amino acids. The methodology is applied to the classification of protein sequences collected from patients with and without amyloidosis, and indicates that the proposed modified classifiers are more robust to sequence variability than standard SVM classifiers, improving classification error between 5 and 25% and sensitivity between 9 and 17%. The classification results might also suggest possible mechanisms for the propensity of immunoglobulin light chains to amyloid formation.

Contact: cyu@bioanalysis.org

*To whom correspondence should be addressed.

1 INTRODUCTION

Critical information relating amino acid changes with the spectrum of functional attributes exhibited by a protein is usually buried among sequence mutations irrelevant for investigated attributes. Immunoglobulin-type beta-domains, which are found in approximately 400 functional distinct forms in humans alone, provide the immense genetic variation within limited conformational changes. A protein database compiled from patients with and without amyloidosis provides unique features to serve as a model system, not only for conformational disease studies but also for the development of computational methods for analysis of structure–function relationships among evolutionarily related families. We are developing computational tools based on the support vector machine (SVM) (Vapnik, 1998) algorithm to classify proteins into pathogenic and benign classes and to identify amino acid variations that contribute to the functional attribute of pathogenic self-assembly in some human antibody light chains produced by patients with amyloidosis.

SVMs have been used recently in a wide variety of applications in computational biology (Noble, 2004). Most applications of the SVM algorithm for protein classification are based on sequence information alone (Jaakkola *et al*., 2000; Hua and Sun, 2001; Leslie *et al*., 2002; Cai *et al*., 2003), as protein structures are usually unknown. Earlier, we developed an iterative SVM-based algorithm for immunoglobulin light chain classification based on protein sequence information (Zavaljevski *et al*., 2002), where each amino acid in the sequence was represented by the same numerical value of six physicochemical properties, regardless of the amino acid position. Identification of the most discriminative sequence positions was accomplished in an iterative procedure by computing a normalized sensitivity index based on the output of the SVM. This approach was successfully applied to a study of the κ_1 subgroup of immunoglobulin light chains and revealed specific amino acid positions in which mutation clearly indicated amyloid formation. However, no similar positions could be found for subgroups of the λ family. This

Report Documentation Page

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
15 JAN 2005		**00-00-2005 to 00-00-2005**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
Classifying noisy protein sequence data: a case study of immunoglobulin light chains	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **US Army Medical Research and Material Command,504 Scott Street,Fort Detrick,MD,21702**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Summary: The classification of protein sequences obtained from patients with various immunoglobulin-related conformational diseases may provide insight into structural correlates of pathogenicity. However, clinical data are very sparse and, in the case of antibody-related proteins, the collected sequences have large variability with only a small subset of variations relevant to the protein pathogenicity (function). On this basis, these sequences represent a model system for development of strategies to recognize the small subset of function-determining variations among the much larger number of primary structure diversifications introduced during evolution. Under such conditions, most protein classification algorithms have limited accuracy. To address this problem, we propose a support vector machine (SVM)-based classifier that combines sequence and 3D structural averaging information. Each amino acid in the sequence is represented by a set of six physicochemical properties: hydrophobicity, hydrophilicity volume, surface area, bulkiness and refractivity. Each position in the sequence is described by the properties of the amino acid at that position and the properties of its neighbors in 3D space or in the sequence. A structure template is selected to determine neighbors in 3D space and a window size is used to determine the neighbors in the sequence. The test data consist of 209 proteins of human antibody immunoglobulin light chains, each represented by aligned sequences of 120 amino acids. The methodology is applied to the classification of protein sequences collected from patients with and without amyloidosis, and indicates that the proposed modified classifiers are more robust to sequence variability than standard SVM classifiers, improving classification error between 5 and 25% and sensitivity between 9 and 17%. The classification results might also suggest possible mechanisms for the propensity of immunoglobulin light chains to amyloid formation.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT **unclassified**	b. ABSTRACT **unclassified**	c. THIS PAGE **unclassified**	**Same as Report (SAR)**	**7**	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std Z39-18

could be explained by the absence of significant single point mutations in this family and/or by a higher degree of sequence heterogeneity in the available data.

Propensity of some proteins to amyloid formation could be characterized by specific sequence motifs, as recently investigated in some experimental studies (Lopez de la Paz and Serrano, 2004). In addition, more genetic variability is present among the λ light chains than among the κ light chains (Williams *et al*., 1996). To enable the analysis of multiple consecutive mutations and account for the high degree of sequence variability, we perform classification based on positional neighborhoods where both sequential and structural neighbors are considered, separately.

Practical assumptions are made in considering structural neighborhoods. Although there are a large number of immunoglobulin structures in the PDB, the vast majority of them are for mice—not humans—and the detailed structural neighborhoods are not known for most of the light chains in our database. However, since immunoglobulin light chains share a similar 3D structure, we assume that the structure of one light chain can be used for the classification of closely related light chains. We anticipate that classification could be improved, in the future, by combining information from molecular dynamics simulations with that of experimentally determined structures to infer structural information that is optimized for each sequence.

2 APPROACH

2.1 Datasets

We use a database of human light chain sequences from 209 patients with and without amyloidosis. Many of these sequences are reported in a previous paper (Stevens *et al*., 1998), and others are available in flatfiles at ftp://bioinformatics.anl.gov/VL-Database/. The database includes both κ and λ gene families encoded on separate chromosomes incorporating substantial amino acid variation. The κ family is represented by four major subgroups, of which the κ_1 subgroup is the most common. The λ family is represented by six subgroups, of which three subgroups are analyzed in this paper. The sequences are manually aligned to 120 positions, taking into account conserved positions in immunoglobulin light chain structures. The variability of the sequences in the analyzed dataset can be quantified by similarity scores based on any established scoring matrix. Table 1 presents the similarity scores based on the BLOSUM80 matrix (Henikoff and Henikoff, 1992). In Table 1, $S(b, p)$ denotes the average similarity score between a chain of class b and a chain of class p, with b denoting the benign class and p the pathogenic class.

Two facts important for the classification accuracy can be observed from Table 1. First, the general sequence variability, described by the standard deviation of the similarity score, shown inside the parenthesis, is much larger for the λ family. This means that there is more noise in the λ family than in the κ family. Second, for each family and each subgroup, there are negligible differences between the average intraclass similarity scores, $S(b, b)$ and $S(p, p)$, and the interclass similarity scores, $S(b, p)$, which represents a problem for sequence encoding based on the amino acid alphabet alone. This implies that a successful classifier ought to use additional information, such as that contained in 3D and sequence structural neighborhoods, so that the encoding (i.e. the weight) of each residue in the sequence is based not only on the amino acid type but on its position in the sequence.

Table 1. Data similarity scores (mean values)

Subgroup	λ_1	λ_2	λ_3	κ_1
Size	28/21[a]	19/20	20/31	36/34
$S(b, b)$	363(173)[b]	427(65)	376(90)	474(43)
$S(p, p)$	419(116)	416(90)	402(90)	467(31)
$S(b, p)$	385(154)	416(81)	387(93)	468(37)

[a]The number of sequences in the pathogenic and the benign classes.
[b]The number in parenthesis represents the standard deviation of the score.

2.2 SVM encoding strategy

Since experimental studies have indicated significant correlation between protein physicochemical and structural properties and protein structural stability (Gromiha *et al*., 1999; Raffen *et al*., 1999), we implement sequence encoding based on six physicochemical properties: hydrophobicity, hydrophilicity, volume, surface area, bulkiness and refractivity (Lohman *et al*., 1994). This type of encoding, therefore, provides additional information for amyloid and benign protein discrimination.

Hence, the encoding of the protein sequence into the SVM algorithm is represented by a real-value vector of dimensionality equal to the length of the protein sequence (120) multiplied by the number of physicochemical properties (6) used to represent each residue. This method enables the SVM kernel function to account for the physicochemical changes in the protein sequences and simplifies the incorporation of the neighborhood information in the SVM algorithm. It is important to point out, however, that while the selected set of physicochemical properties used here was proven to be successful in our previous work (Zavaljevski *et al*., 2002), it is, most likely, not the optimum set. Many different physicochemical encoding strategies could be used; and the identification of the near-optimum set and its implications on the classifier are a worthwhile effort to be considered.

2.3 Methods

The classification is based on a similarity measure of a protein sequence with protein sequences of known attributes. Thus, the similarity measure, also known as kernel function, plays an important role in the SVM algorithm. In addition to standard

SVM kernels, such as the linear kernel (LK), the Gaussian kernel and the polynomial kernel, a variety of string kernels, such as the mismatch kernels (Leslie *et al*., 2002), have been designed specially for protein and gene classification. The mismatch kernels are based on inexact-matching occurrences of k-length subsequences ($k$mers).

Here, we extend a standard kernel that takes the inner product of two vectors representing two protein sequences to kernels, that first average the properties of a residue and its sequential or geometrical neighbors for each residue of the sequence and then take the inner product of the two vectors with averaged property entries. This allows for an area-to-area comparison instead of a position-to-position comparison conducted by a standard kernel. The position-to-position comparison, as the simplest representation, is able to discriminate amyloidogenic proteins characterized by point mutation. Taking into account environmental structural changes in a neighborhood area, the area-to-area comparison should be more suitable to discriminate amyloidogenic and non-amyloidogenic proteins characterized by multiple mutational and genetic differences in the amino acid sequence.

Two kernels are introduced in this paper, sequential and structural (geometric) kernels. The geometric kernel, denoted as GeoNB, is defined as

$$\begin{aligned} K(\boldsymbol{x}_k, \boldsymbol{x}_m) &= K(S(\boldsymbol{x}_k, T), S(\boldsymbol{x}_m, T)) \\ &= K(\boldsymbol{s}_k, \boldsymbol{s}_m), \end{aligned} \tag{1}$$

where $\boldsymbol{x}_k$ and $\boldsymbol{x}_m$ are vectors representing two amino acid sequences k and m respectively. S denotes a mapping from the linear sequence to the 3D structure, and T is the threshold that limits the size of the 3D neighborhood to be considered. We use the maximum neighbor distance suggested in protein structural studies, i.e. $T = 8.0$ Å (Gromiha *et al*., 1999). The elements of the vectors $\boldsymbol{s}_k$ and $\boldsymbol{s}_m$ are represented by weighted averaging of the physicochemical properties in the geometric neighborhood. The average value of property j for position p in sequence m is denoted by $s_{m,p,j}$ and given by

$$s_{m,p,j} = \frac{\sum_{i=1}^{n_p} x_{m,I_p(i),j} w_{I_p(i)}}{\sum_{i=1}^{n_p} w_{I_p(i)}}, \tag{2}$$

where $\boldsymbol{I}_p$ is the vector of neighbor positions for position p, $x_{m,I_p(i),j}$ is the value of property j for the residue at position $I_p(i)$ in sequence m, n_p is the number of neighbors at position p and $w_{I_p(i)}$ is the weight for the neighbor with index i in the vector $\boldsymbol{I}_p$. The weights are defined as the difference between the maximum distance T and the neighbor distance. The distances between two residues are computed using the Contact of Structural Units software (Sobolev *et al*., 1999).

Since the exact structures for each sequence in our data set are not known, we use a representative structural template. A different structural template is used for each immunoglobulin light chain subgroup to determine the 3D neighbors for each position in the amino acid sequence. It is assumed that the geometrical neighborhoods are conserved, i.e. the neighbor positions and their distances for each sequence in the database are the same as those of the template. This assumption could be lifted in the future through the use of molecular dynamics refinement algorithms for the template structural information.

The second kernel, denoted as SeqNB, is the sequential kernel. This kernel is also described by Equations (1) and (2), but the number of neighbors around the residue, designated by n, is specified *a priori* along the sequence. A fixed average distance $\Delta = 1.3$ Å between any two consecutive residues is assumed and used to compute the weights $w_{I_p(i)}$. The distance between two residues separated by i positions in the sequence is $i\Delta$. For symmetric neighborhoods with $n/2$ neighbors on each side, the threshold T for the weight computation is $n\Delta/2$.

Note that Equation (2) explicitly defines the feature space and that the kernel in Equation (1) is computed as the inner product of these features. As a consequence, the Mercer condition (Vapnik, 1998) is satisfied and these kernels are valid kernels.

This classification problem is run on a previously developed computer program, ActiveSVM (Yu and Zavaljevski, 2003), which employs an efficient implementation of the active set method for solving the quadratic optimization, along with two regularization parameters to provide control for the sensitivity and specificity of the classifier (Veropoulos *et al*., 1999).

3 RESULTS

3.1 Classification performance

The ActiveSVM algorithm with three different kernels was applied to four subgroups of immunoglobulin light chains. The geometric kernel is denoted by GeoNB(id), where id represents the PDB identification of the selected template. The sequential kernel is denoted by SeqNB(n), where n represents the number of sequential neighbors in the sequence segment of length $n+1$, with $n/2$ neighbors on each side. The third kernel in our implementation is LK. The LK is selected here to represent a standard kernel, as it was found to be the best kernel in our previous study (Zavaljevski *et al*., 2002). Table 2 shows the performance based on the leave-one-out training/testing procedure. In addition to the overall classification error, Table 2 also presents the classifier sensitivity. For this application, sensitivity is considered more important than specificity. The standard SVM classification with pure sequence encoding, i.e. without property averaging, in the LK is compared against the geometric kernel and the sequential kernel. The sequential kernel calculations are performed for two sequential neighborhoods of lengths $n = 2$ and $n = 4$. The geometric kernel calculations are performed for four different immunoglobulin light chain structural templates for each of the four subgroups, λ_1, λ_2, λ_3 and κ_1. One structural template, identified by its structural database code (1BJM, 1DCL, 1LIL and 1REI), is selected from each immunoglobulin light chain subgroup.

Table 2. Classification performance

Sub group		LK	SeqNB(n)		GeoNB(id)			
			$n=2$	$n=4$	1BJM (λ_1)	1DCL (λ_2)	1LIL (λ_3)	1REI (κ_1)
λ_1	e(%)[a]	33	**22**	22	39	29	29	31
	s(%)	68	**86**	82	68	71	75	71
λ_2	e(%)	44	35	**28**	39	39	41	39
	s(%)	53	63	**74**	58	63	53	63
λ_3	e(%)	35	43	37	35	33	**26**	31
	s(%)	45	45	55	55	55	**75**	55
κ_1	e(%)	**23**	30	26	30	36	30	34
	s(%)	**72**	69	75	69	64	67	64

[a]e: error; s: sensitivity.

The results in Table 2 show significant variability in kernel performance for different subgroups. The best results for each subgroup are highlighted in bold face.

While averaging improves performance for the highly variable λ family, it has a detrimental effect on the κ family. In the previous study, several critical point mutations were found in the κ family. When the sequences that have low noise content are averaged, averaging reduces information content. On the contrary, for sequences with high variability, averaging can improve the signal to noise ratio and thus improves classification. This is the case for the λ family, where averaging consistently provides better performance than the standard LK.

A rather surprising result is the critical dependence of the performance of the geometric kernel on the selection of the structural templates. A significant improvement is obtained for only the λ_3 subgroup. However, it is probable that more specific structural templates could improve the results for the other groups as well. Without a structural template, the classification error for the λ_3 subgroup is 35% while the structural template 1LIL reduces the error to 26% with a significant increase in sensitivity from 45 to 75%. This is the best kernel for the λ_3 subgroup. The performance results using the structural templates from the other immunoglobulin light chains (1BJM, 1DCL and 1REI) are also improved for this subgroup, when compared with the results provided by the LK. The best kernels for subgroups λ_1 and λ_2 are SeqNB(2) and SeqNB(4), respectively, although for λ_1 the difference in the performance between SeqNB(2) and SeqNB(4) is insignificant. The number of neighbors ranging from 2 to 4 is similar to the 3-amino acid motifs found in some amyloid proteins by Lopez de la Paz and Serrano (2004). Small motifs do not appear to be significant for the λ_3 subgroup.

Figure 1 shows the receiver operating characteristic (ROC) curves for each of the three λ subgroups. It shows that the results obtained with the best kernel for a given subgroup is substantially better than those obtained with the LK.

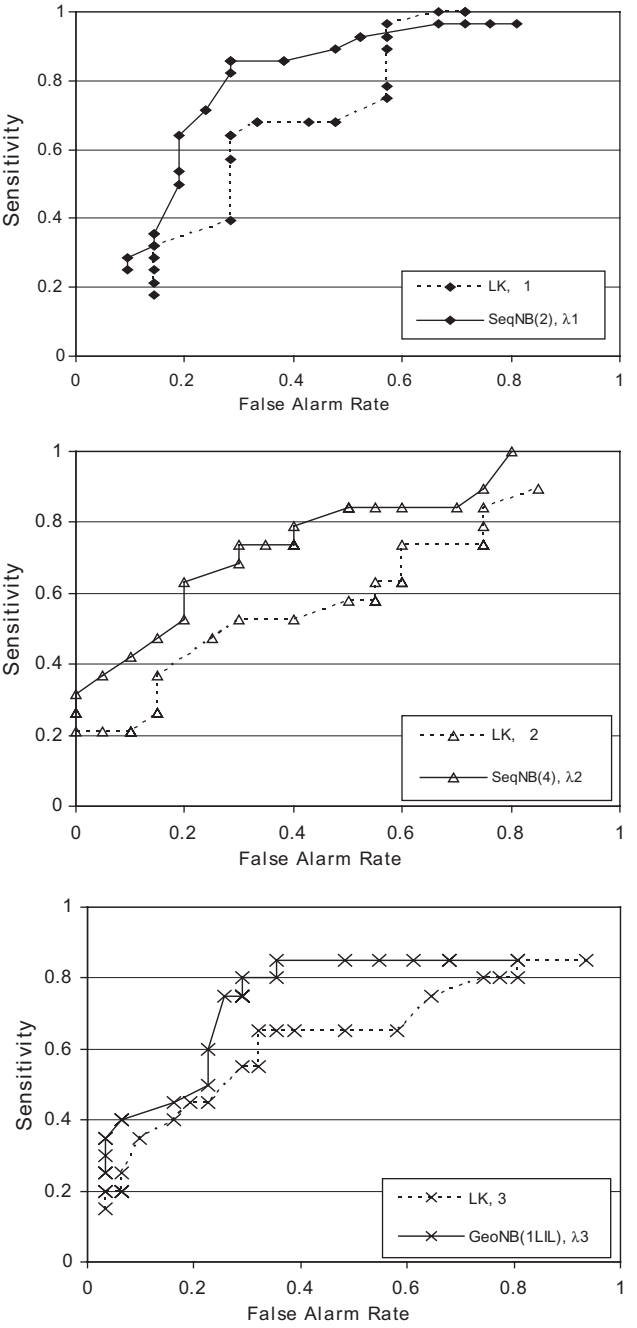

Fig. 1. ROC curve for the λ family.

Since the datasets are very small, the observed improvement in the sequential kernel is tested for statistical significance using resampling. Subgroups λ_1 and λ_3 are resampled by generating 20 random samples from the benign class and 20 random samples from the pathogenic class. The size of the dataset for the λ_2 subgroup being too small for this procedure, this subgroup is resampled together with the λ_1 subgroup to produce a dataset of 35 benign and 35 pathogenic proteins.

Table 3. Error, sensitivity and significance results of the SVM classification

Subgroups	Error (%) Mean LK	SeqNB	Significance (P-value)[a]	Sensitivity (%) Mean LK	SeqNB	Significance (P-value)
λ_1	37.9	28.3	6.0×10^{-11}	58.1	67.3	1.0×10^{-8}
$\lambda_1 + \lambda_2$	37.3	33.1	1.0×10^{-4}	59.8	66.5	9.2×10^{-6}
λ_3	43.3	38.0	4.5×10^{-3}	53.4	62.5	1.5×10^{-5}
$\lambda_1 + \lambda_2 + \lambda_3$	40.2	38.2	5.4×10^{-2}	57.9	63.1	1.5×10^{-4}

[a]The P-value indicates the probability that the differences between two results are due to chance.

Finally, all data are pooled together to produce a dataset of 45 benign and 45 pathogenic proteins. Averaging is performed using $n = 4$ neighbors for the λ_1 and λ_2 subgroups and $n = 6$ neighbors for the λ_3 subgroup, as Table 2 and additional simulations (not shown here) suggest a larger neighborhood for λ_3. The average results over 50 such resamplings are given in Table 3. The Wilcoxon signed rank test (Myers and Well, 2003) is performed on the error and sensitivity results for each subgroup. The results show statistically significant improvement in performance when sequential averaging is used in the SVM kernel. Improvement in sensitivity is more significant. The performance for the pooled data is worse than the performance for the individual subgroups and is driven by the larger classification error of the λ_3 data.

3.2 Possible biological interpretations

Classification results suggest that the mechanisms of amyloid generation might be different for the λ_3 subgroup, perhaps related to a difference in intrinsic propensity towards fibril formation.

Further insight into possible mechanisms for this subgroup is gained by calculating the scores χ_p^2 for position p defined as

$$\chi_p^2 = \sum_{j=1}^{6} \sum_{b=1}^{B} \left[\frac{(m_{pbj}^{+} - P_{pbj} m^{+})^2}{P_{pbj} m^{+}} + \frac{(m_{pbj}^{-} P_{pbj} m^{-})^2}{P_{pbj} m^{-}} \right], \tag{3}$$

where j is the index for the residue properties, B is the number of bins used to partition the probability distribution for each property, b is the bin index, m_{pbj}^{+} is the number of pathogenic samples at position p with property j in bin b, m_{pbj}^{-} is the corresponding number of benign samples, m^{+} is the number of pathogenic samples, m^{-} is the number of benign samples and P_{pbj} denotes the probability that property j at position p belongs to bin b. A large score at a specific position indicates increased importance of that position for discriminating between the benign and pathogenic classes.

The effect of the structural template can be analyzed by computing the difference $\boldsymbol{X}_p$ in χ_p^2 scores, defined as

$$\boldsymbol{X}_p = \chi_p^2\Big|_W - \chi_p^2\Big|_T, \tag{4}$$

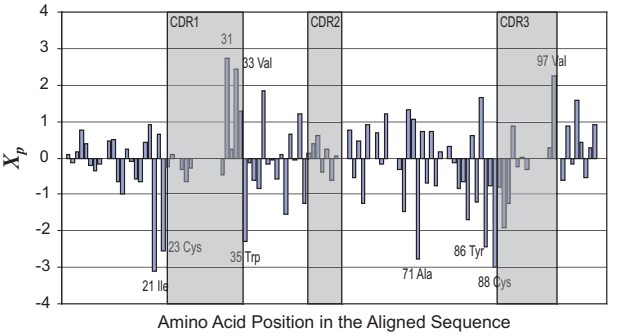

Fig. 2. Difference in χ_p^2 scores for each position without and with the 1LIL template.

where $\chi_p^2|_W$ denotes the score computed without the structural template and $\chi_p^2|_T$ denotes the score computed with the template 1LIL. This difference for the 120 amino acid positions is presented in Figure 2. The three highlighted regions are highly variable regions outside of the protein hydrophobic core, known as the complementarity-determining regions (CDRs). It has been suggested that amyloidosis is related to the protein hydrophobic core (Hoshino *et al*., 2002). As a consequence, CDRs contribute less to amyloid formation. When the structural template is introduced, a significant increase in importance (denoted by a negative value in $\boldsymbol{X}_p$) of some positions outside of the CDRs can be observed. The importance of the variable regions is either suppressed or insignificant, except for a few positions in CDR3. The overall effect of the structural template improves amyloid discrimination, since the importance of regions that are expected to contribute to amyloid formation, such as hydrophobic regions, is increased with respect to less informative CDRs. The significance of hydrophobic regions has also been observed by Serrano's group (Lopez de la Paz and Serrano, 2004; Fernandez-Escamilla *et al*., 2004).

The difference between benign and pathogenic proteins around positions 23 and 88, where two cysteines make a disulfide bridge to stabilize the proteins 3D structure, might have some biological significance. Although the bridge is conserved, the residues around the cysteines are different for benign and pathogenic proteins and might lead to decreased

stability and amyloidosis. The difference at the position of Tyr86 also has structural importance. This amino acid is involved in the classic 'tyrosine corner' in which it forms a buried hydrogen bond to the backbone carbonyl of Asp82. The salt bridge between Asp82 and Arg61 was implicated as a 'risk factor' in κ family amyloidogenesis (Stevens, 2000).

4 CONCLUSIONS

Preliminary results presented in this study indicate that modifications of the standard SVM kernels improve discrimination of benign and pathogenic sequences in the presence of high sequence variability. Proper neighborhood structures are applied for averaging of physicochemical properties that encode sequence data. Thus, the major contribution of this work is the provision of an encoding strategy, which together with special kernel functions tailored for this application provides a mechanism for differential weighting of each residue in the sequence that considers the interactions with neighboring residues. In this way, the encoding of each residue in the sequence considers not only the amino acid type in that position but also the location of the amino acid in the sequence.

For the specific case of immunoglobulin light chains, the variability of neighborhood structures among light chain subgroups might suggest various mechanisms of amyloid formation for each subgroup. For example, for the κ_1 subgroup, propensity for amyloid formation could be traced to single point mutations at specific positions. For the λ_1 and λ_2 subgroups, short motifs of 3–5 amino acids in protein sequences could indicate propensity for amyloid formation. Interpretation of mechanisms for the λ_3 subgroup is more difficult, but might suggest effects of non-local interactions in amyloid formation, since a significant improvement is obtained for this subgroup only when structural neighborhood is included. However, due to very limited data, these conclusions are only tentative and should be validated as more experimental data become available. The importance of a larger database of human immunoglobulin light chains is particularly critical for determination of risk factors in the form of single point mutations or sequence motifs. For larger datasets, more sophisticated methods for sequence motif extraction could be implemented.

Feature selection, i.e. the identification of the key amino acids in a sequence that are important in the characterization of protein function, is of great importance in the development of protein classifiers. It reduces the dimensionality of the input space while reducing the amount of noise in the data (Zavaljevski *et al.*, 2002). Hence, future efforts will be devoted to identifying new feature selection strategies. In particular, we shall further investigate the recently developed statistical mechanics algorithm TANGO (Fernandez-Escamilla *et al.*, 2004). Preliminary evaluation on several proteins in our database indicates that TANGO scores could be used to eliminate non-informative regions in protein sequences prior to classification. In this manner, we could reduce the dimensionality of the encoding vector input to the SVM, reducing noise and potentially improving classification accuracy.

Another future direction for potentially improving protein classification is the computation of optimized structural templates. Strategies to be evaluated could include: creating models that incorporate all (human and non-human) sequences in the database and employing molecular dynamics for protein structure refinement. A second strategy addresses missing templates, i.e. germline representatives for which no structural representative currently exists in the database. In this case, models would be constructed by amino acid replacements of the most similar representative in the database, followed by energy minimization/molecular dynamics.

Many functionally diverse proteins share very similar folds. The distinction between amyloidogenic and non-amyloidogenic proteins is analogous to the distinction of proteins that have known function from those that do not have that function. Increasingly, due to increases in the number of known structures and improvements in recognition of fold at low levels of sequence similarity, it is possible to identify a probable fold. We anticipate that optimized incorporation of structural information with SVM algorithms could contribute significantly to the generation of functional hypotheses for proteins of currently unrecognized function.

ACKNOWLEDGEMENTS

The authors wish to express their gratitude to the anonymous referees for the useful comments and suggestions as well as interesting ideas for future work. The work presented in this paper was supported by the Laboratory Directed Research and Development Program of Argonne National Laboratory and by NIH grants DK43957 and AG18001 (FJS). Argonne National Laboratory is operated by the University of Chicago for the US Department of Energy. J.R. was supported by the Research Area Directorates of the US Army Medical Research and Materiel Command, Ft. Detrick, Maryland.

REFERENCES

Cai,C.Z., Han,L.Y., Ji,Z.L., Chen,X. and Chen,Y.Z. (2003) SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence. *Nucleic Acids Res.*, **31**, 3692–3697.

Fernandez-Escamilla,A.-M., Rousseau,F., Schymkowitz,J. and Serrano,L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.*, **22**, 1302–1306.

Gromiha,M.M., Oobatake,M., Kono,H., Uedara,H. and Sarai,A. (1999) Role of structural and sequence information in the prediction of protein stability changes: comparison between buried and partially buried mutations. *Protein Eng.*, **12**, 549–555.

Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Hua,S. and Sun,Z. (2001) A novel method of protein secondary structure prediction with segment overlap measure: support vector machine approach. *J. Mol. Biol.*, **308**, 397–407.

Hoshino,M., Katou,H., Hagihara,Y., Hasegawa,K., Naiki,H. and Goto,Y. (2002) Mapping the core of the β_2-microglobulin amyloid fibril by H/D exchange. *Nat. Struct. Biol.*, **9**, 332–336.

Jaakkola,T., Diekhans,M. and Haussler,D. (2000) A discriminative framework for detecting remote protein homologies. *J. Comput. Biol.*, **17**, 95–114.

Leslie,C., Eskin,E., Weston,J. and Noble,W. (2002) Mismatch string kernels for discriminative protein classification. *Neural Information Processing Systems 2002*, Vancouver, December 9–14.

Lohman,R., Schneider,G., Nehrens,D. and Wrede,P. (1994) A neural network model for the prediction of membrane-spanning amino acid sequences. *Protein Sci.*, **3**, 1597–1601.

Lopez de la Paz,M. and Serrano,L. (2004) Sequence determinants of amyloid fibril formation. *Proc. Natl Acad. Sci. USA* **101**, 87–92.

Myers,J. and Well,A. (2003) *Research Design and Statistical Analysis*. LEA, Mahwah, NJ.

Noble,W.S. (2004) Support vector machines applications in computational biology. In Schoelkopf,B., Tsuda,K. and Vert,J.-P. (eds), *Kernel Methods in Computational Biology*. MIT Press, Cambridge, MA, pp. 71–92.

Raffen,R., Dieckman,L.J., Szpunar,M., Wunschl,C., Pokkuluri,P.R., Dave,P., Wilkins,S.P., Cai,X., Schiffer,M. and Stevens,F.J. (1999) Physicochemical consequences of amino acid variations that contribute to fibril formation by immunoglobulin light chains. *Protein Sci.*, **8**, 509–517.

Sobolev,V., Sorokine,A., Priulsky,J., Abola,E.E. and Edelman,M. (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatic*s, **15**, 327–332.

Stevens,F.J. (2000) Four structural risk factors identify most fibril-forming kappa light chains. *Amyloid: Int. J. Exp. Clin. Invest.*, **7**, 200–211.

Stevens,F.J., Weiss,D.T. and Solomon,A. (1998) Structural base of light chain-related pathology. In Zanetti,M. and Capra,J.D. (eds), *The Antibodies*, Vol. 5. Harwood Academic Publishers, Australia, pp. 175–208.

Vapnik,V. (1998) *Statistical Learning Theory*. Wiley, New York.

Veropoulos,K., Cristianini,N. and Campbell,C. (1999) Controlling the sensitivity of support sector machines. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI99)*, Stockholm, Sweden.

Williams,S.C., Frippiat,J.-P., Tomlinson,I.M, Ignatovic,O., Lefranc,M.-P. and Winter,G. (1996) Sequence and evolution of the human germline V_λ repertoire. *J. Mol. Biol.*, **264**, 220–232.

Yu,C. and Zavaljevski,N. (2003) *ActiveSVM User's Manual*, Argonne National Laboratory. Argonne, IL.

Zavaljevski,N., Stevens,F.J. and Reifman,J. (2002) Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics*, **18**, 689–696.

DISCLAIMER

The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or as reflecting the views of the U. S. Army or of the U. S. Department of Defense. "This paper has been approved for public release; distribution is unlimited."